Extracting Micro Ontologies from Interrogative Domains For Epistemic Agents

Tracey Hughes and Cameron Hughes

Ctest Laboaratories
One University Plaza
Youngstown, Ohio, 44555
traceyhughes@ctestlabs.org
cameronhughes@computer.org

Abstract

A relevant and functional ontology continues to be one of the bottlenecks to the process of building epistemic agent oriented systems. While the construction of electronic ontologies is the focus of many ongoing efforts, ontology building in a timely manner remains an obstacle. Our current focus is directed toward the notion of automated identification of ontological artifacts from interrogative domains in real time. The artifacts that we are interested in form the basis for a micro ontology of the interrogative domain under consideration.

Introduction

Basic knowledge acquisition can proceed only after the fundamental ontology of the domain has been identified (Nirenburg and Raskin 2004). The problem of identifying an appropriate knowledge representation scheme is also constrained by the selection of a pertinent ontology (Brachman and Levesque 2004). It s for this reason an improvement of the ontology generation process will help to remove one of the bottlenecks to the process of building knowledge-based agent-oriented systems. While there are many efforts underway in the discipline to remedy this (Witherell, et. al. 2010), ontology building remains an obstacle (C-Y. Lu 1987). Our current focus is directed toward the notion of automated identification of ontological artifacts from interrogative domains in real-time. The artifacts that we are interested in form the basis for a micro-ontology of the interrogative domain under consideration. The ontological artifacts of interest are:

- vocabulary of the ontology
- base relationships in the ontology
- primary entities and objects of the ontology

Interrogative Domains as Sources of Knowledge

Interrogative domains consist primarily of questions and answers. A presenter presents an entity with one or more questions and the entity is challenged with providing an appropriate answer (Lehnert 1978). The interrogative domains we are concerned with are digital transcripts of trials, congressional hearings, interviews, surveys and law enforcement interrogations. What we find interesting about transcripts from these areas is the sheer scope of the subject matter. Legal transcripts cover argumentation on such wide ranging topics as when life begins in the womb to violations of civil rights by the patriot act. Transcripts of civil proceedings range from expert testimony on the effect of harmful contaminants in ground water to the effect of video games on adolescent weight gain. It is clear that interrogative domains present particularly fertile sources for knowledge and the value of extracting even micro ontologies from that knowledge should not be underestimated.

The knowledge acquisition process which is dependent on an underlying ontology can be expensive, error prone, and lengthy (C-Y. Lu 1987), and is in fact one of the bottlenecks to building epistemic-based agent systems. Our approach has been to find and experiment with new and different knowledge sources for the knowledge acquisition process (Hughes and Hughes, 2009). The scope and form of digital transcripts make them reasonable candidates for our investigation.

In this paper we describe epistemological and propositional knowledge analysis techniques that we are investigating at Ctest Laboratories. These analysis techniques are used to automatically discover ontological artifacts within the transcripts of an interrogative domain. These ontological artifacts are then used as the fundamental basis for building micro-ontologies of the subject matter found in each transcript. We are using ROGUE (Real-time Ontology Generation Using Epistemic Agents) to perform the transcript and text mining. ROGUE is an experimental multi-epistemic agent-based system under development at Ctest Laboratories. Ultimately, it is ROGUE agents that automatically generate the micro-ontology used as the basis

for knowledge acquisition and a knowledge space for an epistemic agent.

Interrogative Domains and Entailment

In a natural language processing or text mining context, it is not always clear when or how one grammatical item is related to another. It is not always clear when the scope or focus has changed or when new referents have been introduced or old ones dropped. This is part of what makes natural language processing challenging (Barton, et. al. 1987). What sets interrogative transcripts apart from other types of texts and documents is that they consist primarily of question and answer pairs. The relationship between a question and answer is predetermined and clear. The fact that the transcript consists primarily of questions and answers greatly simplifies the segmenting task (Blackburn and Bos 2005). Because questions and answers have a predetermined relationship, we can take advantage of interrogative entailment for transcript mining purposes. Most question and answer pairs entail one or more statements. For example, the question and answer pair in Table 1 is taken from one of the trial transcripts that we used in our data sets.

Question and Answer
Q: *And you saw her take us all to a site that was in London is that correct?* ***A****: Yes.*
Entailed Proposition
E1. *I saw her take us all to a site that was in London.*
Inferred Propositions
I1: She took me to a site in London.
I2: We all went to a site in London.

Table 1. The Q&A pair and there entailed and inferred propositions.

The simple meaning of the Q combined with A semantically entails Proposition E1 and from E1 we may infer Propositions I1 and I2. These propositions compose the propositional knowledge extracted from the digital transcripts.

Notions of Knowledge and Truth

Our discussion of knowledge is restricted to propositional knowledge. We use the Tripartite Analysis (Kant 1965) of Knowledge (TAK) as the basis for our epistemic agents. Although the Tripartite Analysis has short comings and has been thoroughly criticized, see (Gettier 1963) for the basic attack on the Tripartite deconstruction, it is well suited for our implementation of epistemic agents. In this analysis, propositional knowledge is understood as justified true belief. We use Kripke structures (Kripke 1963) as an

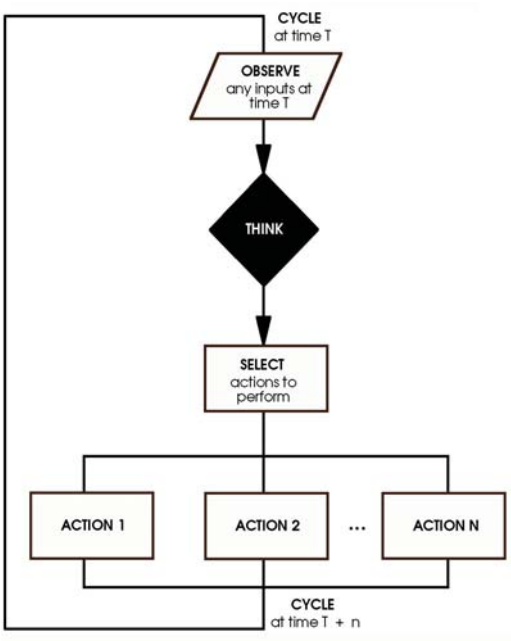

Figure 1. The classic observe think act agent cycle.

intersection between the Tripartite Analysis and our Epistemic Agent. If we let M be a Kripke structure:

$$M = (S, \pi, K_1, \ldots, K_n) \quad (1)$$

Then *K, what the agent knows*, and *S, the worlds the agent consider possible*, are used to represent the notion of modal truth in the context of the Tripartite Analysis. Further, S is expanded and defined by our agents Epistemic Structure (Hughes, et. al. 2008).

Contrast of Agents and Epistemic Agents

In the context of our work and this paper, an agent is recursively defined as a situated autonomous software component that has an agent cycle. A distinguishing feature of the agent is that it affects and is affected by the environment that it is situated in through the activity that takes place in the agent cycle as seen in Figure 1. While there are several classic notions of what constitutes an agent cycle, see (Shoham 1993), (Rao and Georgeff 1995) and (Wooldridge 2000), we take our cue from Kowalski s and Sadri s work in logic programming and multi-agent systems (Kowalski and Sadri 1999). We chose to build on (Kowalski and Sadri 1999) partially because it subsumes the notions described by (Shoham 1993) and (Rao and Georgeff 1995) but primarily because their unified agent cycle is compatible with our notion of an epistemic agent (Hughes and Hughes, 2009). In (Kowalski and Sadri 1999) unified agent cycle, the "*think*" process utilizes reasoning, logic programs, and integrity constraints. The goals of the agents include commands and queries, integrity constraints,

obligations and prohibitions, condition-action rules, and commitment rules.

Epistemic agents have an epistemic structure (Hughes, et. al. 2008), and their agent cycle is focused on acquiring, justifying, and maintaining propositional knowledge and classifying propositions which serve as the goals and the actions that can be performed. Further, epistemic agents have a DNA (Deductive Nuclear Architecture) (Hughes and Hughes, 2009) as opposed to a BDI architecture (Wooldridge 2000). The DNA is a structure designated by α :

$$\alpha = < E_s, \Theta, I_t, \delta > \quad (2)$$

where :

E_s is the Epistemic Structure,
Θ is a deductive theory of inference over E_s
I_t, is a set of interrogative types,
δ is the deductive response, the set of conclusions reachable from E_s.

The epistemic structure is a knowledge representation scheme that models the TAK (justified true belief). It is defined as:

$$E_s = <G_1, G_2, J_s, V_c, F> \quad (3)$$

where:

G_1 is a graph of a *priori* propositions,
G_2 is a graph of a *posteriori* propositions,
J_s is a set of justification propositions,
V_c is a vector of commitment,
F is a non-monotonic truth maintenance function on E_s.

The a priori propositions stored in G_1 constitute the entailed propositions from the Q&A pairs of the digital transcript. The propositions in G_2 are inferred and acquired during the agent cycle and induction. Initially G_2 is { }, and we have G_1 with a cardinality > 0. Some of the a priori propositions will serve as justifications (populate J_s) and as a basis for the agent's level of commitment (populate V_c) (Hughes and Hughes 2009) to the other propositions in G_1 and to those that will later populate G_2 . The agents beliefs are based upon the V_c determined by J_s . *Initially*, the agent evaluates its level of commitment to the proposition contained in G_1 . As propositions are inferred and G_2 becomes populated, the level of commitment to propositions in both G_1 and G_2 are re-evaluated by F which in turns updates J_s and V_c. F functions as an integrity constraint assuring that the agents propositional knowledge is logical and consistent. Populating G_2 is determined by Θ , the deductive theory of inference which defines the inference types. Depending on the proposition type (determined by the Q&A type from which it was entailed based on I_t) will dictate the inference type. Θ *and* I_t both works as a condition-action rules.

Defining the Knowledge Space

Let $d = \{ E_{s1}, E_{s2}, E_{s3}, ..., E_{sn} \}$

where d is a set of epistemic structures representing a particular domain, or a collection of domains. Then we have:

$$K_s = \bigcup_{i=1}^{N} d \quad (4)$$

where K_s is a set union of the domains, or the total knowledge space of the agent, also called the *epistemic reality*. Building K is where one of the primary bottlenecks to deploying epistemic agent-oriented systems. Typically,

Techniques	Description	Pros/Cons
Inte v ews	L ve Q & A sess ons w th actors used to produce protoco They can be structured, sem and unstructured	**Pros:**
		Va dat on s mmed ate
		Cons:
		Exp c t know edge not tac t know edge
Quest onna res	Quest ons presented to a respondent that supp es wr tten answers	**Pros:**
		Standard stat st ca treatment S mp e to comp ete Not t me consum ng
		Cons:
		Answers are m ted L tt e nteract on
Obse vat on	Observ ng the expert perfo m ng tasks and tak ng notes n order to document protoco s	**Pros:**
		Low comp ex ty and cost Gather process and tac t know edge
		Cons:
		T me consum ng process Tota dependency on obse ver
So t ng	Used to capture the way respondents compare and order concepts, objects, attr butes, and va ues assoc ated w th them	**Pros:**
		E c t deta know edge about concepts
		Cons:
		Concepts must f rst be dent f ed Experts may have va y ng v ews

Table 2. Description, pros and cons of KA techniques.

K is built using one or more of several standard knowledge acquisition techniques. These techniques are shown in Table 2.

It is in the knowledge acquisition phase where we look to speed up the process. Rather than acquiring the base ontology through the traditional processes of conducting interviews, disseminating questionnaires, repository evaluation local-remote observation and requirement analysis (Whitman, et. al. 2007), we are exploring the use of transcript mining (Hughes and Hughes 2009) to automatically identify the base ontology for the agent's a priori knowledge space. Once it's built, we can summarize the behavior of the epistemic agent.

If we let:

$$e_n \;/\; \alpha \tag{5}$$

be a set of agents with DNA then the Fundamental Epistemic Axiom can be stated as:

$$q\;\delta(q) \;\leftrightarrow\; K_s \quad \delta(q) \quad e_n/\alpha\,\Theta(\,I_t(q)\,)\;\cap\;M \tag{6}$$

This axiom is used to guarantee the epistemological soundness of the work performed by the epistemic agents.

Ontological Artifacts

The ontological artifacts are taken directly from the propositional knowledge found in the transcripts. If we consider the transcript as the universe of discourse, then all the worlds that the agent considers possible will be contained within the transcript since it provides the complete context for the epistemic agent's knowledge. So when it comes to truth analysis, the epistemic reality (4) and the worlds the agent considers possible (Kripke 1963) are taken from the transcript. With this in mind we use transcript mining (Hughes and Hughes 2009) to automatically extract the entailed propositions of the transcript.

There are 5 basic steps in the extraction of the ontological artifacts from the transcript:

Step 1: Segment the transcript into blocks of Q&A pairs while maintaining the chronological appearance of witnesses, attorney's evidence, etc.

Step 2: Classify the Q&A pairs according to the 13 basic categories of questions and answers.

Step 3: Resolve anaphora, and mark substitutions between the question and answer pair blocks.

Step 4: Using interrogative entailment, convert the Q&A pairs to propositions.

Step 5: Using the predicates: `simple_subject()`, `simple_predicate()`, `simple_object()` extract the ontological artifacts from the propositions

The classification of the Q&A pairs has important ramifications because the Q&A pair classification determines the type of entailed proposition. If the Q&A pair is a location type, then the entailed proposition will also be a location type and will make an assertion about location. It is also important to do the Q&A classification at this point because it helps with the anaphora (Kamp and Reyle 1993) resolution rules.

So, if we let: $T = \{$ *set of Q&A in transcript* $\}$ then:

$$T \quad T_l \tag{7}$$

where T_l is the set of propositions entailed from T.

Model Theoretic Semantic = Micro-Ontology

Once we have T_l we perform a Model Theoretic Analysis (MTA) on T_l to extract the ontological artifacts of our interest (Blackburn and Bos 2005), being:

- vocabulary of **T**
- base relationships in **T**
- primary entities and objects in **T**

Note that the MTA gives us the artifacts that form the basis of our micro-ontology. In fact, an MTA is sufficient to produce a micro-ontology for a transcript knowledge source. This analysis is done using three predicates from the ROGUE system:

```
simple_subject(Pi)
simple_predicate(Pi)
simple_object(Pi)
```

where `Pi` is an interrogatively entailed proposition. And the three simple predicates take `Pi` as arguments and produce the corresponding ontological artifact. For example, we can extract ontological artifacts for the entailed proposition in Table 1:

E1: *I saw her take us all to a site that was in London.*

In Step 3, the anaphora are resolved so the propositions will contain these substitutions:

E1 w/ substitutions:
Mr. Garcia saw Ms Runga take Mr. Garcia, Mr. Cannon, and Mr. Homes to a site that was in London.

The resulting extracted ontological artifacts would be:

was_in(Mr. Garcia,London_site)
was_in(Mr. Cannon,London_site)
was_in(Mr. Homes,London_site)

The ontological artifacts are captured by the model theoretic semantics of the transcript. Here the model is described as `m = (D,f)` where `m` is a model theoretic semantic representation of all the language that is contained in the trial corpus. `m` consists of a pair `(D,f)` where `D` is the *Domain* which is the set of people and things referenced in the corpus (e.g. defendants, jurors, attorneys, witnesses) plus the relations (e.g, `lawyer(X,Y)`, `trial_day(N)`, etc.) between those people and things. `f` is an *Interpretation Function* which maps everything in the language onto something in the domain (Blackburn and Bos 2005). `f` is implemented using the Lambda Calculus operator λ and β-conversion. `m` captures completely our ontological artifacts. It is important to note here that the process that extracts `m` from T_1 requires robust natural language processing (NLP). At Ctest Labs, we use ISIS, an agent-oriented system dedicated to NLP. The micro-ontology is a component of the Cognopaedia which serves as the universe of discourse.

ROGUE Agent Architecture

As mentioned earlier at Ctest labs, we are using ROGUE system to extract these ontological artifacts. ROGUE (Real-time Ontology Generation Using Epistemic Agents) is an experimental multi-epistemic agent-based system that automatically generates the micro-ontology used as the basis for knowledge acquisition and a knowledge space for an epistemic agent. The ROGUE agent exists in an environment from which it senses percepts (inputs), perform actions, and then outputs to the environment. There can be 1 to n ROGUE agents in a given environment working in parallel. Figure 2 shows the architecture of the ROGUE agent.

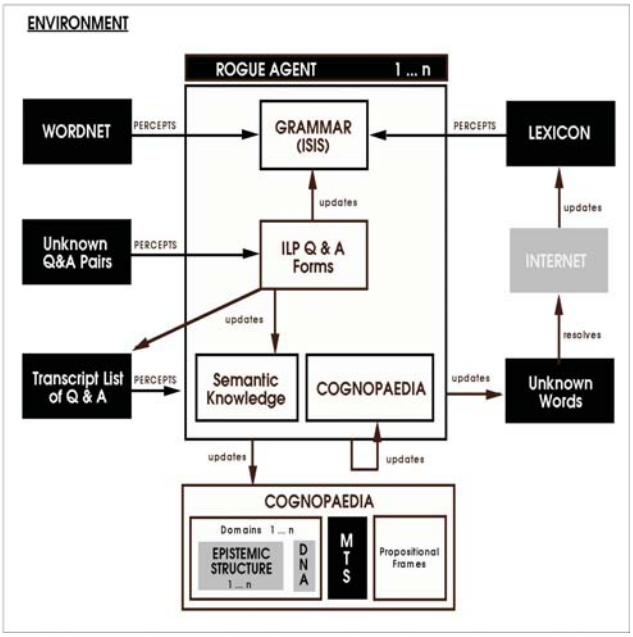

Figure 2. The architecture of the ROGUE Agent.

The ROGUE agent has GRAMMAR (ISIS) and ILP Q&A Forms. ISIS (interrogative Sentence Interface Subsystem) provides a NLI (Natural Language Interface) to the ROGUE system. It is involved in processing the Q&A pairs and mapping them to propositions and OAV tuples. ISIS receives inputs from the WORDNET and Lexicon components. ILP Q&A Forms is an inductive logic program that processes unknown Q&A pairs. When ILP becomes aware of an unknown Q&A form, it attempts to identify it using ILP. If it is able to identify the unknown Q&A form, it updates ISIS so it can now process that form when it is again presented. The now known Q&A pairs of that form is then re-introduced to the agent from the Transcript List Q&A component. The ILP Q&A Forms also updates the agent s Semantic (Procedural) Knowledge.

The COGNOPAEDIA is an ontology and knowledge construct that replaces the application domain. It contains 1 to nE_s, an ontology (and micro-ontology), and propositional frames that models the propositions. When propositions are entailed or inferred, and tuples are extracted, the E_s and ontology contained in the COGNOPAEDIA are updated. The "on board" knowledge construct is required by the ROGUE agent in order to update the COGNOPAEDIA that is apart of the environment. The agent s "on board" knowledge construct is updated by the agent itself.

Discussion And Future Work

In this paper, we have presented some of our work at Ctest Laboratories where we are using ROGUE to extract the micro-ontology or model theoretic semantic from digital transcripts. We are interested in the transcripts because of the sheer scope and subject matter that transcripts cover. The notion that we could improve traditional knowledge acquisition techniques by automating the process of building the micro-ontology is very enticing. However, we may be too ambitious in our objectives because to date, we ve only built micro-ontologies for less than two dozen transcripts and the transcript subject matter was not sufficiently diverse. Also we lacked diversity in the types of transcripts that we used. In particular, we have only looked at trial transcripts as opposed to interviews, congressional hearings, surveys, interrogations, etc. However, because building the micro-ontology is usually a prerequisite for the full knowledge acquisition phase, we are persuaded that the ROGUE approach will prove useful and we are already in the process of setting up more formal and exhaustive experiments. We are in the process of attempting multi-domain knowledge spaces.

Finally, the question and answers themselves require robust natural language processing in the interrogative entailment phase. From the transcripts we have processed, about 20% of the Q&A pairs are still not intelligible to our grammar and semantic parsers. So while the digital transcript as a knowledge source on the surface appears to be very fertile ground, we still have much work to do. We

believe that access to a large number of transcript will help the transcript mining process produce multi-domain models. Currently we are using data sets of less than two dozen transcripts due to the difficulty of access. We believe that the more transcripts we consider on a topic the higher the knowledge credential and payoff.

We are considering ways we can enhance our ISIS NLP system as a result of its failure to process 100% of Q&A pairs presented to it. We still see a great deal of promise in using digital transcripts to automate parts or all of the knowledge acquisition process.

References

Witherell, P., Krishnamurty, S., Grosse, I., and Wileden, J. 2010. Improved Knowledge Management Through First-Order Logic in Engineering Design Ontologies. Journal of Artificial Intelligence for Engineering Design, Analysis and Manufacturing Vol. 24: 245-257.

C-Y. Lu, S. 1987. Knowledge Map: An Approach to Knowledge Acquisition in Developing Engineering Expert Systems. Journal of Engineering with Computers Vol 3: 59-68.

Breitman, K., Casanova, M.a., Truszkowski, W. 2007. Semantic Web Concepts, Technologies and Applications. NASA Monographs in Systems and Software Engineering: Springer.

Hughes, C., and Hughes, T. 2009. Transcript Mining Using Epistemic Agents and Interrogative Entailment. In Proceedings of IEEE International Conference on Intelligent Computing and Intelligent Systems, Vol. 1, 857-861. Beijing, China: IEEE Press.

Kant, I. 1965. The Critique of Pure Reason. trans Norman Kemp Smith. New York: St. Martin.

Gettier, E. 1963. Is Justified True Belief Knowledge? Analysis, Vol.23: 121-123.

Kripke, S. 1963. Semantical Considerations on Modal Logic. Acta Philosophica Fennica, vol. 16: 83-94.

Hughes, T., Hughes, C., and Lazar, A. 2008. Epistemic Structured Representation for Legal Transcript Analysis. Advances in Computer and Information Sciences and Engineering. Springer: 101-107.

Shoham, Y. 1993. Agent Oriented Programming. AI Journal 60(1): 51-92.

Rao, A.S., and Georgeff, M.P. 1995. An Abstract Architecture for Rational Agents. In Proceedings of the International Conference on Logic Programming, 67-81. Kanagawa, Japan: MIT Press.

Wooldridge, M. 2000. Reasoning About Rational Agents. Cambridge, Massachusetts: MIT Press.

Kowalski, R., and Sadri, F. 1999. From Logic Programming towards Multi-Agent Systems. Annals of Mathematics and Artificial Intelligence, Vol 25, issue ¾: 391-419.

Kamp, H., and Reyle, U. 1993. From Discourse to Logic, Introduction to Model Theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory. Netherlands: Kluwer Academic Publishers.

Brachman, R. J., and Levesque, H. J. 2004. Knowledge Representation and Reasoning. San Francisco, CA.: Morgan Kaufman.

Barton, G. E., Berwick, R.C, and Ristad, E.S. 1987. Computational Complexity and Natural Language. Cambridge, Massachusetts: MIT Press.

Blackburn, P., and Bos, J. 2005. Representation and Inference for Natural Language. Stanford, CA.: CSLI Publications.

Nirenburg, S., and Raskin, V. 2004. Ontological Semantics. Cambridge, MA.: MIT Press.

Lehnert, W.G. 1978. The Process of Question Answering. Hillsdale, NJ.: Lawrence Erlbaum Associates, Inc.